Applied Text-Mining Algorithms for Stock Price Prediction Based on Financial News Articles

Adrian Besimi
South East European University, North Macedonia
a.besimi@seeu.edu.mk

Zamir Dika
South East European University, North Macedonia
z.dika@seeu.edu.mk

Visar Shehu
South East European University, North Macedonia
v.shehu@seeu.edu.mk

Mubarek Selimi
South East European University, North Macedonia
ms21693@seeu.edu.mk

This article includes a developed model and well-defined process that one should undertake in order to contribute in the prediction of the potential stock price fluctuation solely based on financial news from relevant sources. We are providing background information on this topic adding the role of text mining in general, furthermore supporting the idea with the study of relevant research articles to narrow the focus on the problem we are researching. Our proposed model relies on existing text-mining techniques used for sentiment analysis, combined with historical data from relevant news sources as well as stock data. In confirming the model, after the experiment we have provided the results of the simulation, which are opening the ground for further explorations in this sensitive area of prediction.

Key Words: text mining, finance, news, crawling, stock, prices, prediction, naíve bayes
JEL Classification: C89, G17
https://doi.org/10.26493/1854-6935.17.335-351

Introduction

The data produced and the speed at which data is provided on the Internet nowadays has increased to a degree and at a rate that is impossible to process. This trend, on the other hand, has challenged the research in many areas, such as data mining and text mining, which are the focus of

our study. These two areas have emerged in the last decade mainly due to research in artificial intelligence, machine learning, and inferential statistics (Vale 2018).

Stock market data and relevant news associated with fin-tech industry are increasing rapidly as well. Many investors that are handling stock market transactions have a major interest in understanding more about the future of stock markets for the purpose of being able to do an educated guess and/or predict any future investment. Ensuring some level of prediction in market fluctuation can assist investors in the form of decision support systems and integrated with existing automatic trader agents that would ensure better prediction on future trades. Fully predicting the market fluctuation means in practice becoming a billionaire over the night and all the time minimizing financial losses, which is not possible for many reasons. Recent scholars argue that news articles are among influential sources that may affect stock market prices and they should be carefully considered by investors when planning future investments. By definition, any stock price is simply defined by supply and demand of the market, but it is argued by scholars Nikfarjam, Emadzadeh, and Muthaiyah (2010) and Kaya and Karsligil (2010) that another important variable when decision is made to invest or not is also related to verifiable news from financial news sources. This is hard and time-consuming task because it requires to read and analyze a lot of news published on several occasions by various news sources/providers (Nikfarjam, Emadzadeh, and Muthaiyah, 2010; Kaya and Karsligil 2010).

Information published in news articles influence, to a varying degree is influencing the decision of the stock traders, especially if the given information is unexpected. It is important to analyze this information as fast as possible, so it can be used as an advantage to help traders to make trading decisions before the market has had time to adjust itself to the new information (Aase 2011).

One important application of using text mining is text sentiment analysis, also known as opinion mining, a technique that digs deep into the content of the text file and extracts the sentiment of it. Sentiment analysis classifies textual data into positive texts, negative text and neutral text sentiments which is later used for the purpose of categorizing any text documents into the given sentiment (Aase 2011; Khedr and Yaseen 2017).

The model proposed in this paper is going to leverage the Naïve Bayesian classifier for document classification to make a prediction for whether the stocks will go up or down, based on a dataset that is generated from the process proposed later in this paper.

Literature Review

Several scholars, specifically Kim, Jeong, and Ghani (2014) prove to some extent in their work that the relevant news are closely related to stock price movements in the market. With the current trends in big data and content creation on the Internet and the enormous amount of unstructured text data available, the mobile channels, and Social Network services, scholars have attempted to predict stock movements using such text data as in the case of Kim, Jeong, and Ghani in 2014.

Many scholars tried different approaches in research to prove that there is a potentially strong correlation among financial news articles and stock price fluctuations, as is the case of Khedr and Yaseen (2017) that we mentioned earlier. In their paper they propose an approach to use sentiment analysis in financial news, along with features extracted from historical stock prices to apply prediction for the future behaviour of stocks. According to their findings, the proposed model has achieved high accuracy using sentiment analysis in categorizing news polarities by applying Naïve Bayes algorithm. In their case the accuracy of the model is up to 86.21%. By moving on with their experiment in prediction, during their next attempt in analyzing these news articles, they have included numerical attributes which in their case increased the accuracy to 89.80%.

The paper, published by Hagenau, Liebmann and Neumann (2013), examines the hypothesis if any stock price prediction based on textual content from the financial news can be further improved. In this paper, the authors have upgraded the text mining methods by adding expressive feature to represent the text and by adding more variables, such as employing market feedback in the feature selection process. According to the authors, this selection of the features does significantly improve the accuracy due to the fact that this approach removes the unnecessary so called 'less-explanatory features,' i.e., noise, which itself helps the classifier to overcome the over-fitting during classification of the text. In the case when the feedback-based feature selection is combined with 2-word combinations, the authors results show an accuracy of 76%. These results are different from common sentiment analysis approaches since the 2-words combination gives more information and potentially more meaning to the sentiment classification.

A lot of research has been carried by scholars in the area of prediction of stocks as well. A project by Joshi, Rao, and Bharathi (2016) is taking financial news articles about a given company, and they use these data to try to predict the future movements of the stock again by applying sen-

timent analysis. The approach is like in the other cases with an idea to identify how stocks have reacted if news has polarity. Authors in this case have taken the past three years of data from Apple Inc. stock prices as well as news articles. Similarly, to previous scholars, the polarity of the news is labelling these articles and based on these data they are building the training set. The approach in this paper is dictionary-based that contains for positive and negative words that is build based on financially specific words. Further, they have pre-processed the data which resulted in having their own finance specific stop words and dictionary. Using their own dictionary, they have implemented three models for classification and tested them. After comparing the results, they have concluded that Random Forest algorithm resulted in better accuracy for the test cases ranging from 88% to 92%. This algorithm was followed by Support Vector Machines with again very good accuracy of 86%. In their case the Naïve Bayes algorithm performance was the lowest with 83%.

There is some promising research published that applies deep learning techniques and has resulted with higher accuracy ratings. Of interest is the published paper by Tabari et al. (2019) that shows a comparison of diverse algorithms specifically applied in stock market tweets. This research shows quite promising results, with accuracy ranging up to 92.7 % (using Convolutional Neural Networks). However, even though deep learning approaches can be scaled for using news articles, other authors report much lower accuracy rates Kim and Jeong (2019).

One major advantage of using Naive Bayes algorithms is its well-known ability to improve by introducing new data. In our case, previously analyzed news articles can be fed to the algorithm and treated as prior probability. With this, new, previously unknown words will gain weight and affect prediction when encountered in the future. This is one major drawback of the proposed model from Joshi, Rao, and Bharathi, mentioned above, since it only works for a pre-defined dictionary of words.

Another similar approach of finding the correlation amongst the content of news articles and stock prices for the purpose of predicting the stock markets was implemented by Kaya and Karsligil (2010). They collected news articles published in the last year period and combined with the stock prices for the same period. These articles were then labelled as positive or negative sentiments categorization based on their effects on stock price. Their approach is a little bit different in the sense that for them it was important to use the price changes for categorization of the news. While analyzing the textual data, authors use and approach of word

doubles of a noun and a verb as features and not only single words. The support vector machines (SVM) method was used in this case which resulted with 61% accuracy.

These scholars and articles mentioned in the section above are the core of our model and study that we conducted.

Methodology

PROBLEM DEFINITION

Financial analysts that are handling investments and transactions in stock markets around the world have a huge headache on making decisions that will be effective and bring more money to the investor or maximizing profits by trading. They are aware that any news, either good or bad can directly affect the stock market. The job of these experts relies on analyzing everything from the media outlet. This is time consuming and the amount of data is getting larger all the time. The methodology that we are arguing as many other scholars mentioned above, including also Falinouss (2017), is that an advanced text mining algorithm can assist these experts and provide them with knowledge just by processing resources related to text and news.

The price movements from the past are not always a good indicator on the future movements and are not a guarantee of smart investment, which makes news articles analysis a better predictor on stock market movements. Falinouss in 2007 proposed to research about the impact of textual data in predicting the financial markets movements. In his study he also developed a system which uses similar approaches as previous case of text mining techniques and their influence on the stock market. This according to Falinouss (2007) can help financial analysts to act immediately upon new news articles as they get published.

We propose a model of predicting stock price fluctuations or movements by analyzing financial news articles on one hand and historical stock prices on the other hand. To accomplish this objective, a complete process of data mining and text mining was developed to predict the price movements for the 3 companies listed public, which are explained in the subsection below.

PROPOSED MODEL FOR STOCK PREDICTIONS BASED ON FINANCIAL NEWS

In our study we worked towards analyzing data, concretely news articles and historical stock prices to make future predictions about stock direc-

tion. To achieve this, qualitative and quantitative data are crucial. Many steps are conducted to achieve the aim of this research, starting from data gathering.

The data is collected for a period of one year, starting from the 1st of March 2018 until 1st of March 2019. In order to make the prediction we used different variables, such as the polarity of the news (either positive or negative), the rate of change in stocks quotes (an average of 5 days), a source of the news article as well as the company name.

The following is the process consisted of eight steps needed to be performed in order to predict stock price from the financial news:

1. Identifying the news sources and targeted companies
2. Data collection and data cleaning of news articles
3. Sentiment Analysis of news articles
4. Data collection of stock prices
5. Calculating Rate of Change (ROC)
6. Categorizing the data
7. Applying Naive Bayesian classifier
8. Training

Identifying the news sources and targeted companies is crucial to understand the data. The information collected must be relevant and trustworthy. As such, the relevant data from financial news articles from top reliable sources have been identified as: The Washington Post, CNN, MarketWatch, BGR, Fox Business, The Street, The Verge and Breitbart. The targeted companies for our study are: Tesla, Facebook and Apple. News sources are proven to be reliable in the market as the most unbiased, whether the targeted companies have been chosen randomly from technology, software and automotive industry. Tesla has been added because it is a typical example of a lot of news noise and several fluctuations of stock prices.

Data collection and data cleaning of news articles. Links of the news from the sources mentioned in step 1 are collected using Web Scraper extension of Google Chrome browser. After having all the links, we built a python script based on Scrapy (an open source and collaborative framework for extracting the data you need from websites, see http://www.scrapy.org). framework that is extracting data from the links and organizing them in the following structure: article's Title, date, author and the text content. Appropriate data cleansing has been applied

to remove unnecessary HTML tags as well as to format the data from different sources to one standard (see table 1 and table 2).

Sentiment analysis of news articles was applied to every news record based on the news content by using Vader Sentiment Analysis. VADER (Valence Aware Dictionary for sEntiment Reasoning) is a pre-built sentiment analysis model included in the NLTK (Natural Language Toolkit) package. It can give both positive/negative (polarity) as well as the strength of the emotion (intensity) of a text. VADER however is focused on social media and short texts, unlike Financial News which are almost the opposite. We updated the VADER lexicon with words plus sentiments from other sources/lexicons such as the Loughran-McDonald Financial Sentiment Word Lists, to be appropriate for our collected financial news (Yip 2018). At the end of this step we had the polarity of the news content recorded in our dataset.

Data collection of stock prices for each of the targeted companies was done from Yahoo Finance portal, where the following information was collected: *date, open price, high, low, close price, volume* and *Adj close.* These data are important for correlation with the appropriate news from our first data set.

Calculating Rate of Change (ROC). The *ROC* and *Future ROC* are the two variables that are calculated from the data set from Step 4. The rate of change (ROC) in stocks in an average of 5 days is an existing formula that refers to the last 5 days of stock fluctuation. In our case we also added a column with the Future ROC (the ROC after 5 days), having in mind that the effect of this positive or negative news will be reflected in the future and not the past. Since we are dealing with historical data, the Future ROC is easy to calculate.

Categorizing the data must be done in order to apply Naive Bayesian classifier. In the data set that we have all news collected with their features, we added two new columns: *Sentimentof_text* that could be 'positive' if the sentiment score is greater than zero and 'negative' if the sentiment score is less than zero. We don't take in consideration the neutral score of the text content because that could result in majority of neutral results. The second column is the *ROC_Sentiment* that can be 'positive' if the Future ROC is greater than zero and 'negative' if the future ROC is less than zero.

Naive Bayesian classifier was used to make the prediction of the future stock movements. The naive Bayes applies the well-known Bayes' theorem, where by using a 'naive' assumption that any set of features are

independent for a given class (Tang, Bo; Kay, Steven; He, Haibo; 2016). To prepare the data set to make predictions with the NB, we added a new column with the name class that is 'UP' if the *Sentimentof_text* is 'positive' and the *ROC_Sentiment* is 'positive,' and if the *Sentimentof_text* is 'negative' and the *ROC_Sentiment* is 'negative' then the class is *'down,'* otherwise is *'neutral'* classification. The training dataset results are summarized in table 3.

Training. The data that is collected (see table 1 and table 2) contains records for 12 months from which 10 months will be used to train the model and the last 2 months will be used for the test set, to evaluate how it performs. In total 18236 records will be used as training dataset and the remaining 1990 records (roughly 10%) out of 20226 will be used as a test set.

We created 2 models to see how they perform. The following variables are used to train and test the first model: *Source, Company, Sentimentof_text* and the *5-day ROC,* while in the second model only variables of: *Source, Company* and *Sentimentof_text.*

Results

As explained in the steps undertaken to perform our prediction, the data collection results are shown below. We succeeded to collect the news articles from 8 different news sources, totalling 20226 news articles, split into table 1 for Training Set (18236 records/articles) and table 2 for Test Set (1990 records/articles).

The training dataset results are summarized in table 3, where for each company in our target the list of classification results is shown. As a general finding is that the algorithm applied classifies 15.71% of the articles in the training set as *'down'* (meaning the stock will go down in the following days), 50.71% is classified as *'neutral'* (there is no clear picture on what the prediction will be) and 33.59% of the data as *'up'* (meaning the stock will go up). The *'up'* classification is relevant to our study and can be used for simulating investments on our test data from the test set.

In the first prediction that uses the following variables: *Source, Company, Sentimentof_text* and the *5-day ROC* model, the test set classification from 1900 records being tested, resulted in 564 *'down'* and 1426 *'up'* classes for stock price direction were predicted. The achieved accuracy of 94.29% in this prediction model shows that there is a very high chance to predict the stock price movements. By this, our arguing that based on several attributes from new articles we can reach a certain level of predic-

TABLE 1 Total News Articles Obtained for Apple, Tesla and Facebook Organized by Source For Training Set (March 2018–December 2018)

Variable	Categories	Frequencies	Percentage
Source	BGR	1073	5.884
	Breitbart	435	2.385
	CNN	687	3.767
	Fox Business	813	4.458
	The Street	3810	20.893
	The Verge	2847	15.612
	The Washington Post	6051	33.182
	Market-Watch	2520	13.819
Company	Apple	7591	41.626
	Facebook	7513	41.199
	Tesla	3132	17.175

TABLE 2 Total News Articles Obtained for Apple, Tesla and Facebook Organized by Source for Test Set (January 2019–March 2019)

Variable	Categories	Frequencies	Percentage
Source	BGR	185	9.30
	Breitbart	167	8.39
	CNN	211	10.60
	Fox Business	147	7.39
	The Street	590	29.65
	The Verge	603	30.30
	The Washington Post	87	4.37
	Market-Watch	0	0
Company	Apple	1144	57.49
	Facebook	416	20.90
	Tesla	430	21.61

tion, and give directions to financial experts, is valid as in our case where we reached a certain level of prediction based on several attributes. Still, though Efficient Market Hypothesis (EMH) clearly states that financial stock prices cannot be predicted, because there is no 100% prediction. The accuracy rate of the first model is high and there is a strong relationship between financial news and stock price movements.

TABLE 3 Training Set Classification Data Organized by Company and Frequency

Company		Down	Neutral	Up	Total
Apple	*N*	1,006	3,930	2,655	7,591
	%	5.52	21.55	14.56	41.63
Facebook	*N*	1,390	3,683	2,440	7,513
	%	7.62	20.20	13.38	41.20
Tesla	*N*	468	1,634	1,030	3,132
	%	2.57	8.96	5.65	17.17
Total	*N*	2,864	9,247	6,125	18,236
	%	15.71	50.71	33.59	100.00

To test the second model 3 variables as input are given: *Source, Company* and *Sentimentof_text* to predict the class up, down or neutral. In comparison with the first model that has an accuracy of 94.29%, the second model has 49.49% which is significantly with lower accuracy than the first model, that has just one more variable – the *5-day* ROC. This prediction rate is less than the guessing probability (50%), and as such this model is irrelevant. It can be stated that aside from sentiment of text, stock data variables as in this case of '*5-day* ROC' plays a vital role in prediction of the future stock price movements. Only a single relevant variable as the rate of change of a given stock for the last five days does impact the prediction in the model where we deal with stock rates. It is important to include the most relevant variables when applying prediction in the FinTech industry.

It can be stated that the weak form of EMH is true. Weak form efficiency states that all future price movements follow a random walk, unless there is some change in some fundamental information. It does not state that prices adjust immediately in the advent of new fundamental information, which means that some forms of fundamental analysis and news article analysis might provide excess returns. This is because they trade on new information and does not use any historical information to look for patterns.

Simulation

In order to apply and validate the text-mining algorithms and classification techniques mentioned in this paper to predict the financial news, we have conducted a simulation based on exact data from the market and

the cross reference of the dates with the predictive algorithm. Simulation data is conducted with 10.000$ investments per company per day.

Table 4 contains sample data for our simulation. Bellow is the description of the fields in order to better understand when the 'investment' is to be made:

- *DoW* represents the Day of the week that we have news articles processed and categorized the sentiment. It is an important variable to understand if we are dealing with weekends.
- *Date* is the actual date when the news have appeared and the stock Close price date.
- *Company* is company name that we have collected data.
- *Close* is the closing price of the stocks for the given date and company.
- *5-day ROC* is the historical rate of change for the last 5 days.
- *Invest?* is a probability calculation that is derived as an average of all sentiments for all news in our dataset. For instance: *if there is one news article and it is positive towards Facebook, then this percentage is 100%. In case if there is one negative and 3 positive news for a given company, then the probability is 75% to invest.*
- *Investment* is the same amount used for the purpose of simplification and simulation: $10,000.00.
- *Profit/Loss* is calculated based on the prediction if we go for investment. For instance, *the second row results in investment for Facebook. In this case 'investment' probability is 100% and in the simulation we 'invest' total $10,000. The difference in Close price for Facebook for 1/1/2019 and Facebook for 1/2/2019 in stock prices is ($10.000/135.68 = 76.283 stocks, or equal on 2nd of January as 76.283x$135.68 = $10,350.14) and profit/loss is $350.14.*

Rules and assumptions for 'investment' are the following:

- The 'investment' is done a day after the news have been published (the effect of the published news will be seen the next day). In case of weekends, the investment is done next Monday.
- The purchased stocks will be sold the next day, for simplification the cost of closed stock price for that day has been taken into the simulation.
- The difference represents either Profit/Loss.

TABLE 4 Profit and Loss Sample Table for the Test Set simulation

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
3	1/1/2019	Apple	157.740	18.423429	80.00	10,000.00	11.41
3	1/1/2019	Facebook	131.089	−60.73621	100.00	10,000.00	350.14
3	1/1/2019	Tesla	332.799	110.98007	100.00	10,000.00	(681.49)
4	1/2/2019	Apple	157.919	20.466856	67.74	10,000.00	(996.07)
4	1/2/2019	Facebook	135.679	−59.230769	66.67	10,000.00	(290.39)
4	1/2/2019	Tesla	310.119	96.601993	57.14	10,000.00	(314.72)
5	1/3/2019	Apple	142.190	8.4674699	54.93	10,000.00	426.89
5	1/3/2019	Facebook	131.740	−60.414660	62.50	10,000.00	471.38
5	1/3/2019	Tesla	300.359	90.197561	54.55	10,000.00	576.97

NOTES Column headings are as follows: (1) DOW, (2) date, (3) company, (4) close, (5) 5-day ROC, (6) invest (%), (7) investment ($), profit/loss ($).

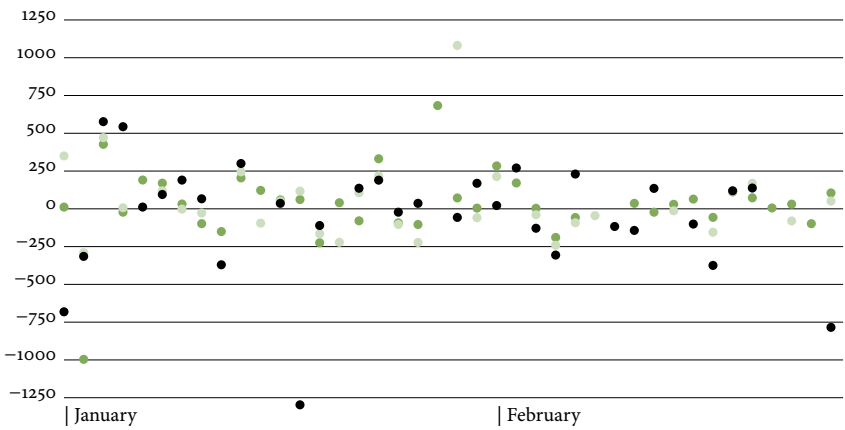

FIGURE 1 Chart of Profit/Loss simulation on Test Set Data Based on Classification Model (Naive Bayes) Used in This Study (dark green – Apple, light green – Facebook, black – Tesla)

Figure 1 clearly shows the fluctuation on the investment simulation based on real data from stock closing prices, for the three companies in combination with our model as explained in previous table. *Tesla* data shows more fluctuation and as such we excluded in our next chart (figure 2) to see if the model prediction can be used for investment.

Figure 2 shows improvements and majority of profit cases on the investment simulation when *Tesla* is excluded, and the only trade is done with *Apple* and *Facebook*. In order to support this figure, table 5 represents

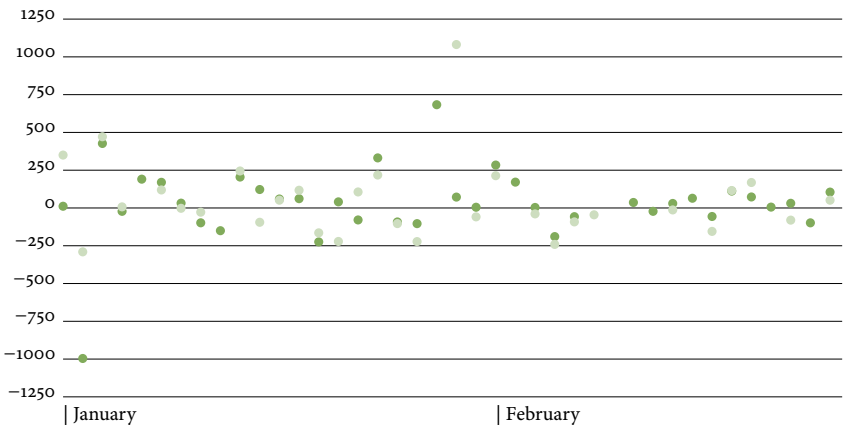

FIGURE 2 Chart of Profit/Loss Simulation on Test Set Data Based on Classification Model (Naive Bayes) Used in This Study, Excluding Tesla Company (dark green – Apple, light green – Facebook

real data for our simulation for the 2 months test data used in this experiment. The values show daily profit/loss for all three companies based on the news sentiments that we processed.

In table 5, Profit and Loss table for Apple, Facebook and Tesla based on the Test Set simulation with $10.000 investments was conducted. The simulation conducted does not show 100%-win case for the classification of stock prediction and as such it does not apply to all companies. The difference where there are better results relies on the targeted companies, such as Apple and Facebook, which are more stable ones rather than Tesla, which as a case had different fluctuations that in long term did not bring good results in our simulation.

In this simulation, Tesla's predictions based on our model result in Loss where the other two companies Apple and Facebook in the long run result in profit of around $2.600 if the algorithm is run every day where the data is available with daily amounts of $10.000 per company investments.

Conclusion

The trading of stock in public companies is an important part of the economy, so in this study stocks have been analyzed through using data mining and text mining techniques to make a prediction for stock price directions of the stocks for 3 companies listed public.

To achieve a prediction we gathered data, collected relevant financial news articles from reliable sources with both qualitative and quantitative

TABLE 5 Profit and Loss Table for Apple, Facebook and Tesla Based on the Test Set Simulation

Date	Apple	Facebook	Tesla	Grand total ($)
1 January	11.411	350.141	−681.490	−319.939
2 January	−996.074	−290.388	−314.717	−1601.179
3 January	426.893	471.382	576.975	1475.250
6 January	−22.258	7.249	543.611	528.602
7 January	190.631		11.644	202.275
8 January	169.817	119.273	94.826	383.916
9 January	31.962	−2.080	190.234	220.116
10 January	−98.180	−27.739	66.383	−59.536
13 January	−150.371		−370.328	−520.699
14 January	204.667	244.859	299.940	749.466
15 January	122.166	−94.663		27.503
16 January	59.378	51.512	36.411	147.301
17 January	61.594	117.329	−1297.112	−1118.189
21 January	−224.461	−164.622	−110.501	−499.584
22 January	40.443	−221.590		−181.147
23 January	−79.262	106.029	136.306	163.073
24 January	331.369	218.062	189.702	739.132
27 January	−92.545	−103.348	−22.219	−218.113
28 January	−103.647	−222.418	36.439	−289.626
29 January	683.347			683.347
30 January	72.012	1081.638	−56.676	1096.974
31 January	4.807	−58.791	169.044	115.060

Continued on the next page

data. This combined with the second type of data of stock prices were used in our study. For every article, a sentiment score (positive and negative) of the text content is calculated.

We have found out that a model that does not include price fluctuations and wholly relies on text content to predict the stock price fluctuation is not accurate at all. Including additional variables improves significantly the prediction. In our case the variable '5-day ROC' plays an important role in predicting the future stock prices.

This article, except for proposing the model used and the process undertaken to arrive at the desired data set, contains results from the sim-

TABLE 5 *Continued from the previous page*

Date	Apple	Facebook	Tesla	Grand total ($)
3 February	284.050	213.626	21.781	519.456
4 February	171.094		270.382	441.477
5 February	3.445	−39.145	−128.520	−164.220
6 February	−189.394	−241.070	−306.096	−736.560
10 February	−57.509	−92.034	230.216	80.673
11 February		−45.238		−45.238
12 February			−116.737	−116.737
13 February	36.433		−142.779	−106.347
14 February	−22.249		135.300	113.052
18 February	29.926	−12.924		17.002
19 February	64.354		−100.773	−36.419
20 February	−56.386	−155.020	−374.471	−585.876
21 February	111.657	115.596	119.492	346.746
24 February	72.845	168.633	137.762	379.240
25 February	5.740			5.740
26 February	30.975	−80.424		−49.449
27 February	−98.359			−98.359
28 February	105.112	51.409	−784.356	−627.836
Grand Total	1135.432	1465.244	−1540.327	1060.350

NOTES Empty fields in the rows above are possible due to the fact that not all companies chosen for investment every day, based on the news sentiment probability of investment.

ulation of the model. Previous models for sentiment analysis of financial news articles are limited in news articles from relevant sources and as such, based only on sentiment of the news do not provide enough information for future movements. Our model in this paper adds more variables to the dataset in order to give more accuracy to the prediction.

As the results are probabilistic weights (predictions), the simulation we conducted does not show 100%-win case for the classification of stock prediction and as such it does not apply to all companies. The difference where we have better results relies on the targeted companies, such as Apple and Facebook, which are more stable ones rather than Tesla, which as a case had different fluctuations that in long term did not bring good results in our simulation.

Future work would follow with the research on the characteristics of

the companies that would fit to the model, with the tendency to prove that the proposed model is universal for the specific companies within specific variables, adding more tests and simulations as well.

It could also prove valuable to evaluate deep learning algorithms for the purpose of sentiment analysis. These algorithms are yet to show good results when larger text bodies are used, however, for short tweets they are very accurate. Furthermore, in this paper we evaluated only a three-class problem in the context of sentiment analysis. It would be of interest to approach using a multiclass prediction model and see how diverse sentiments would affect the stock market. This is an area where NN algorithms would prove to be much more beneficiary. Finally, even though we consider only highly respected news sources for our analysis, we could further drill down to add the author of the source as an attribute. A renowned author would probably have more weight in the context of affecting the stock market with his articles.

References

Aase, K. G. 2011, 'Text Mining of News Articles for stock Price Prediction.' Master's thesis, Institutt for datateknikk og informasjonsvitenskap.

Falinouss, P. 2007. 'Stock Trend Preidction Using News Articles: A Text Mining Approach.' Master's thesis, Luleå University of Techonology.

Hagenau, M., M. Liebmann, and D. Neumann. 2013. 'Automated News Reading: Stock Prices Prediction Based on Financial News Using Context-Capturing Features.' *Decision Support Systems* 55 (3): 685–97.

Joshi, K., J. Rao, and H. N. Bharathi. 2016. 'Stock Trend Prediction Using News Sentiment Analysis.' *International Journal of Computer Science & Information Technology* 8 (3): 67–76.

Kaya, M. Y., and M. E. Karsligil. 2010. 'Stock price Prediction Using Financial News Article.' In *Proceedings of the 2nd International Conference on Information and Financial Engineering,* 478–82. Chongqing: IEEE.

Khedr, A. E., and N. Yaseen. 2017. 'Predicting Stock Market Behavior Using Data Mining Technique and News Sentiment Analysis.' *International Journal of Intelligent Systems and Applications* 9 (7): 22–30.

Kim, H., and Y.-S. Jeong. 2019. 'Sentiment Classification Using Convolutional Neural Networks.' *Applied Sciences* 9 (11): 2347. https://doi.org/10.3390/app9112347

Kim, Y., S. R. Jeong, and I. Ghani. 2014. 'Text Opinion Mining to Analyze News for Stock Market Prediction.' *International Journal of Advances in Soft Computing and its Application* 6 (1): 1–13.

Nikfarjam, A., E. Emadzadeh, and S. Muthaiyah. 2010. 'Text Mining Approaches for Stock Market Prediction.' In *Proceedings of the 2nd Inter-*

national Conference on Computer and Automation Engineering, 256–60. Singapore: IEEE.

Tabari, N., A. Seyeditabari, T. Peddi, M. Hadzikadic, and W. Zadrozny. 2019. 'A Comparison of Neural Network Methods for Accurate Sentiment Analysis of Stock Market Tweets.' In *ECML PKDD 2018 Workshops,* edited by C. Alzate, A. Monreale, L. Bioglio, V. Bitetta, I. Bordino, G. Caldarelli, A. Ferretti, R. Guidotti, F. Gullo, S. Pascolutti, R. G. Pensa, C. Robardet, and T. Squartin. Cham: Springer.

Vale, M. N. d. 2018. 'Dow Jones Index Change Prediction Using Text Mining.' Instituto Alberto Luiz Coimbra De Pós-Graduação E Pesquisa De Engenharia, Rio de Janeiro.

Yip, J. (2018), 'Algorithmic Trading Using Sentiment Analysis on News Articles.' https://towardsdatascience.com/https-towardsdatascience-com-algorithmic-trading-using-sentiment-analysis-on-news-articles-83db77966704.